>> Greg Pewett: Welcome and thank you for standing by. At this time, all participants are in a listen-only mode. Today's webinar is being recorded and will be posted publicly. If you have any objections, you may disconnect at this time. Now I'd like to turn the call over to Meghan Maury. Meghan, you may begin when ready.

>> Meghan Maury: Thank you so much and hello. I'm Meghan Maury. I'm a senior advisor here at the Census Bureau. And I'm so excited to welcome you to today's Data Toolbox: Administrative Data panel. We really appreciate you spending some time with us today so we can share with you a conversation about how administrative data are used in the Census Bureau's work. Now if you've been with us throughout the series to date, you've heard a great administrative data 101, I hope, and then an intriguing conversation about how administrative data are used in really innovative and impactful work we do across the Census Bureau. In today's session, we're going to keep rounding out your knowledge by exploring two of the products the Census Bureau is best known for, the decennial census and the annual population estimates. They'll tell some similar information about the people in the United States: total population, characteristics like age, sex, race, and ethnicity; the population living in group quarters; number of housing units. But those two products are created in really different ways, and some of the ways that they're different really are rooted in administrative data. So as most of you know, the decennial census is primarily created using direct responses from households, but administrative data are used to supplement those responses where we have, for example, missing or incomplete data. And it's used in a couple of other ways in the decennial census as well. Population estimates, on the other hand, use several sources of administrative data as really the primary tools to estimate the population, how it changes, and those direct responses from households are just one part of the input to what we call the population estimates base. So in this conversation, we'll explore some of the similarities and differences and how those two data products are developed and how administrative data play a role in increasing the accuracy of both products. But first I'm really excited to introduce today's panelists. So Tom Mule will speak to us first about the decennial census. Tom is the special assistant to the chief of the Decennial Statistical Studies Division in the Census Bureau. He's on the administrative records modeling team for the 2020 Census that researched and implemented the usage of administrative records and third-party data to reduce the number of inputs and visits we had to make when we had really high-quality administrative records, so he's deeply knowledgeable about all of this work. Eric Jensen similarly is a senior technical expert for the demographic analysis in the Census Bureau's Population Division, which means he leads the demographic analysis work, which you'll hear about later, and he works on population estimates as well. He also works really extensively on the undercount of young children. Some of you may have heard him talking about that topic as well. Welcome Tom and Eric. So glad to have you here. Now let's quickly make sure that all of our listeners are acclimated to Webex, which may be a new platform for some, so that they can really easily ask questions during the panel today. So to ask a question, you'll want to look for the Q&A feature, and note that this is not the chat function. The chat function is something separate that we'll use to communicate with you if need be. But to make sure that we see your question, please make sure you're putting it in the Q&A feature. You'll see that hopefully at the bottom right-hand side of your screen. If you don't see it there, you may need to click on

three little dots at the bottom right and then you'll see an option to open the Q&A feature. But you may see a small box with a question mark in it. Let us know if you're having a hard time finding those. Now, when you type your question, please choose the option that lets you send it to all panelists to make sure that someone here on our team sees it. And once our conversation winds down, we'll grab some of the questions you've sent in and we'll talk through the answers here with Tom and Eric. So, lastly, this webinar is being recorded, as we said at the beginning, so that we can make it available to you on our website for continued reference. And I think that's all of our logistical pieces. So let's jump in. Let's start with the decennial census. So, you know, although 2020 seems like so long ago already, we're actually still releasing data from that census. There were a few different ways we used administrative data in the 2020 Census, and even when we did use administrative data, that wasn't always as simple as counting a whole household with one single administrative record. It was much complicated than that. And I'm hoping that you can walk us through that a little bit, maybe through an example of what might have happened with a hypothetical household.

>> Tom Mule: Yeah. Thank you, Meghan. I'd be happy to do that. A lot of times when people are thinking about how we use administrative data in the census, they're thinking about one component of the 2020 Census, and that's where we used administrative record data when a whole household hadn't responded to the census. So as part of today, I would like to provide some more nuances about how we did that in the 2020 Census and plus also like to provide some additional context on some different ways that administrative data was used. Next slide please. So today I'd like to talk about the part of the decennial census use of administrative data during our data collection and processing phase. So today I'd like to be able to spend some time going over three areas where we were able to do this during the 2020 Census. So one of them was when we were able to identify vacant or nonexistent addresses, those that did not meet our definition of a housing unit. Secondly, we'd like to expand a little bit more on what were we able to do to use administrative record data for occupied units. And then, third, we'd also like to be able to talk a little bit about how were we able to use administrative data for the assignment of demographic and housing characteristics. If you were able to join us last week during Meghan's session, she went a little bit over the edit and imputation, and so I'd like to just provide a little bit more information about that today. So next slide please. So first we're going to start off with the first bucket and then also during the 2020 Census when we had nonresponding addresses, the Census Bureau needed to determine whether the address met one of three categories. So first of all was it occupied, was it an address where people are living there, what was the roster, and what was the characteristics. The second outcome that we needed to determine was, okay, if people are not living there, okay, that's a vacant address. So we needed to determine that and also try to determine what was the reason why the address was vacant on Census Day. And the third was that we had addresses on our list that were not housing units. Those might have been businesses. Those also might have been no longer habitable. So, first of all, I'd like to talk about the first bucket, but how can we identify addresses that were vacant or they did not meet our definition of being a housing unit. So for the 2020 Census, we did a lot of research leading up to the 2020 Census and being able to see how can we use information that we already had available to us to be able to identify that an address was vacant or that it was not a housing unit. And so in the course of us doing that,

what we'd like to mention is some of the information that we were able to use to make those determinations. So, first of all, one of the main pieces of information was that when postal carriers were delivering our census mailings around Census Day, the postal carriers, just like when you received your mail at your house or apartment today, the postal carriers were able to let us know whether they were able to deliver our census mailing of the letters and the questionnaires to the address or not. If they weren't able to deliver it, the postal carriers were able to let us know that that was undeliverable as addressed. In addition, they were able to provide us the reasons why they could not deliver it, so they were able to indicate that they could not deliver because the address was vacant or there was no such number or there was no such street, so other reasons as why they could not deliver the census mailing to that address. We also have other information from our administrative data about that address. So we also had information about whether the people at the address, at the household, did they file their tax returns? Did that household, were they able to receive their W2 information earlier that year? We have also information is anybody at that household, are they participating in Medicare? Are they a patient in the Indian Health Service? We also have information from a third-party source source are they seeing anybody living at the address? And also in course of doing this, we're also able to take into account information from our American Community Survey. From our American Community Survey, we have information and estimates about how often do we think people in that area, are they moving, vacancy rates, able to take inside account. So based on that information, we're able to make initial determination if we think that address is vacant or if we think that address is not a definition of our housing unit. During the 2020 Census for those identified addresses based on that information that we had around Census Day, we proceeded to also send out another mailing to the address. So in 2020 in June, about eight weeks after Census Day, we proceeded to send another postcard to that address. In the course of doing that, the postal carrier was able to determine whether they were able to deliver that piece of mail again eight weeks later. So based on that and also as part of our 2018 test, while we had tested during the decade conducting no visits to these addresses, we decided after our 2018 end-to-end test to visit these addresses at least once and having a measure at least of being able to make sure that we were doing additional measures to make sure that we were counting everyone. As part of doing this one visit, if we had the initial determination around Census Day and then also unable to deliver around that June delivery, those were candidates for this one visit. And also as part of all this was going on while some of these addresses may have only received one visit during our nonresponse followup operation, our internet response was available, so households could respond online throughout October.

>> Meghan Maury: That's so neat. And I really love that example because it just makes you rethink, you know, what constitutes an administrative record. Like I think of the postcards we sent out to households and I often am thinking through like what was the content of that postcard. Was it motivating for the recipient household? But we're also just getting information about whether or not it was delivered, and that's another data point we can use in our process. I really love that. And it's really, you know, an example too of just a different way that we use administrative records that wasn't about, you know, counting people or filling in information but more about using administrative records actually to make sure we're not missing people. Can you expand on that a little bit? I mean, are there

other ways we're using administrative data to help us know whether we're counting everybody or counting all the people at an address?

>> Tom Mule: Yeah, I'd be happy to do so. So next slide please. As part of our innovation area for the 2020 Census, we also focused on it to say how could we use this administrative data that we had available to identify the people living at an address and the characteristics. So we built on lessons based on our research and the testing that we were able to do in the 2013, 2014, 2015, 2016 census tests and plus also our 2018 end-to-end test. So based on all these tests, we were able to go through and able to identify how we can roster people from what taxes, how were they able to file based on their -- filing their taxes at the address, being able to roster people based on the W2 or 1099 information that we had for the address. We were able to roster people from the Medicare enrollment database. We were also able to take into account are people associated with that address because they are participating in the -- a patient in the Indian Health Service. We were also able to take advantage of files that we make at the Census Bureau to be able to see, okay, can we link children to their parents. So we might have an infant where a W2 form might have the parents and we're able to use this other information to link the children to those parents to make sure we're trying to achieve a complete household enumeration. We're able to look and see do we have multiple sources that are indicating that the family lives at that address, and in the course of doing that, we're doing the testing throughout the decade, we were able to refine our procedures. And our procedures, we were able to come up with ways that can address you know, couple of questions. One, are we counting all the people that we've rostered from these different sources in the right place. We don't want to be counting some of the people if we're going to be reducing contact for an address. We want to make sure we're counting all the people and in the course of doing that making sure we're doing everything we can that we are not missing people. One of the results from our 2020 analysis is that for taking our administrative records rosters, when we compared them to addresses where we got a self-response and that self-response came in before the end of July, we were able to compare our administrative records rosters to that actual census count that came in. And one of the things that we were happy to see is that for these households, our adrec (administrative record) data rosters, they agreed with the census count 83.8% of the time. So we were happy to see that agreement.

>> Meghan Maury: Yeah, that's so interesting. I can't wait to read more about that. You know, I actually have a question for Tom, and the question that comes to me is like, let's say I'm a person who lives in a home that's really impacted by the pandemic or let's say a natural disaster and I wanted to respond to the census but I wasn't able to prioritize it. I didn't get to it. And the census takers keep coming to my house, but, you know, I'm at work. I'm missing them. But I do show up in administrative data. Can you walk us through like how does the census approach trying to count me, especially given that, you know, when there is that good administrative data available?

>> Tom Mule: Yeah. I'd be happy to do so. So with the next slide -- -- so this gives an overview using our self-response type of enumeration area. This is the numerous ways that we attempted to contact households during the 2020 Census to encourage them to respond. So the first line shows plans for the 2020 Census because we had numerous mailings during our self-response period during March and April where we were sending

multiple letters and/or, you know, questionnaires to the address encouraging them either to respond online or using the paper questionnaire that was sent along the way. While we may have determined that we had good administrative record information, still during our nonresponse followup operation, we did do one visit at the door. So we were able to go to the household. So the interviewer in the course of knocking on the door, one possibility is they can complete the enumeration with the householder during that visit. Another possibility is that during that visit the enumerator could determine that address was vacant or it wasn't a housing unit. So we could get that correct information right there on the doorstep. But also if they weren't able to get a resolution during that contact, we were able to leave a notice of visit on the door. So this did allow the household, when they did end up coming home from work, they could see the notice of visit, peel it off the door, and this was another way of prompting them to go online and fill out their questionnaire or being able to pick up their paper questionnaire and, you know, send it back in. Additionally, about one week after that visit happened, we sent another postcard to the address encouraging them to respond. In addition, the self-response internet option was open all throughout the nonresponse followup operation, so people could have responded all the way through October in 2020. So there were these multiple ways that we were encouraging people to respond to the census where we were going to use the administrative record information that we had.

>> Meghan Maury: Interesting. So let's say I did fill out the census but I didn't fill it out completely, like maybe I just wrote down the people living in the house but not the characteristic information like their age or their race or ethnicity. How did we use administrative data in that context?

>> Tom Mule: Yeah. So this is one [inaudible] definitely be able to share and highlight another usage that people may not have been aware. So in the 2020 Census, we were able to use administrative data in two ways. The first was if we did need to use our administrative data for the roster for the households that we built, if we needed to use that for enumeration, we could use the administrative record data information that we had to fill out the characteristics of that household. So that was one usage. Another usage is that people may have self-responded or responded to one of our enumerators during the census but they may have not provided an answer or left a question blank. And this was another instance where we were able to use administrative record data to help fill in those blanks. So with the administrative record data, for this discussion, I'm going to expand it to also include past census data but also responses that we may have received from the American Community Survey. So with us using past census responses, the 2020 Census was not the first time this was done. In the 2010 Census, the Census Bureau used race and Hispanic origin responses that were provided through the 2000 Census. They were able to use those in the 2010 Census if we didn't have that information. But for our 2020 Census, we expanded on this to say, okay, how can we build on this and use other information. So we ended up looking to say, okay, how can we use information from our Social Security Administration. They're able to indicate if a person moved into the country, what country did they move from? We were also able to look and see for other programs like Medicare or Indian Health Service. With participating in those programs, you're able to indicate your race and Hispanic origin. So we were able to identify how we could use information that

was gathered when participating in other programs and being able to take advantage of that. So next slide please. So another area where we're able to use administrative record data was for age information. So with age information, we were able to use information from our past census. We were also able to use information that was collected during the American Community Survey. We were also able to take into account was their information that was available from the Social Security Administration. When a child is born and is applying for their Social Security Administration number, the Census Bureau is able to get that -- we're able to get that information delivered to the Census Bureau and use that. And similarly for sex. Sex is another instance where we have that information from the past census but also Social Security Administration when applying with birth certificates and being able to get your Social Security number is able to capture that information and deliver to us as well. So with using age as an example, next slide please. We were able to take an instance where we have the Doe family where they were able -- we had John Doe and Baby Doe reporting at 101 Main Street, and John Doe did not provide their age. So we're able to look either at past census or Social Security Administration information to identify that, oh, John Doe was born on February 2nd, 1969 and be able to account for that information and be able to assign that instead of having to impute it. Next slide please. Another characteristic that we had I explained before was race and Hispanic origin, so we're trying to be able to capture the race information because I explained we're able to build off our 2010 usages. So an example for like Hispanic origin, if somebody was able to respond to the 2010 Census that they were Puerto Rican, then we were able to use that detailed information in the 2020 Census. Next slide please. With the racial information, an example is that we might have in this instance John Doe replying to the census but not providing their race. And if John Doe moved to this country from Japan and Social Security Administration was able to let us know that, we would be able to take advantage of that information, be able to assign that John Doe was Asian and plus also be able to capture the more detailed information that John Doe was Japanese. And next slide please. So, yeah, so with building off of this work that we were able to do using administrative data for race and Hispanic origin, we were able to take a selection of our 2020 Census people and be able to link them to the administrative record data that we had available. So with this, we were able to break the people down into 14 categories based on whether they were Hispanic or non-Hispanic but also their race, whether each one of our race alone, including some other race or also multiracial. So we were able to take our responses, able to link those to our administrative record data, and we're able to see in that instance that when we could do that link, that 89% of the time our census reporting in the 2020 Census matched what we had available in administrative record data. So there was another very good amount of agreement that we were able to see but also with going forward for the 2030 Census how can we continue to work on this to address the 11%, especially with multiracial populations. So these are just some examples of how we were able to use administrative data in the decennial census and also we'll talk about that in more detail during the course of us having this session today. So I hope this was able to help folks understand these three major contexts of where we were able to use administrative data in the 2020 Census.

>> Meghan Maury: Thank you so much. It's so fascinating, you know, and as I learn more about this, I'm just eager to learn more. And I loved on your first couple of slides where you were showing that this is even just one piece of the process. We even use administrative

records in building the frame and in other ways in the decennial census too that we didn't even get a chance to touch on today. So really, really interesting. Now, you know, most people know that the census is used for a lot of sort of really big purposes like congressional apportionment and redistricting. You may also know that census results are used to allocate, you know, hundreds of billions of dollars in federal funds and as population controls for surveys and as denominators for calculating vital rates and other statistics. But the census is something we only take once a decade, so for the other years, we need to rely on something else to make sure that we're using updated information. And for many of these purposes, we actually rely on the official population estimates, these kinds of things. So every year we produce a new timeseries of the population estimates that start at the date of the most recent census and extend to the vintage year which represents the last year of the data available. And to do this, we use something that we refer to as the population balancing equation, which we shorthand as the new estimates equals the base population plus birth minus death plus or minus net migration. And I think we have that equation on a slide here somewhere, but we'll get to it. So next we're going to take a deep dive into how administrative data or data from other federal, state, and local programs are used to produce these population estimates. So, Eric, I hope you can come join us on the stage, the proverbial stage. And I think a good place to start might be to talk about the historical context of using administrative data for these purposes. My hunch is that the conversation around administrative data can sometimes feel a little new to folks, but when it comes to population estimates, I mean, we've used these for a really long time, right, and those estimates are really grounded in administrative data. So it's not a new topic to you.

>> Eric Jensen: That's right, Meghan. Next slide please. So as Tom mentioned, the decennial census relies mostly on self-response and administrative data are used when we're not able to get that self-response. The population estimates program is different in that we have a long history of using administrative data. In fact, we've always relied on vital records as the main source for input for the estimates. By vital records, I mean records that are produced when births and deaths are registered, and this is done when a person fills out a birth or death certificate. We don't just use administrative data for births and deaths though. We also use administrative data to estimate both international migration and domestic migration. So now I'm going to walk through the different components of population change and focus on the administrative data that we use to produce each one. So let's start with births. Next slide please. So currently we use microrecords on births from the National Center for Health Statistics to produce the estimates of children born each year. Again, these records are created when the birth certificate is filled out. In the United States, birth registration is more or less 100% complete in that practically all births are registered.

>> Meghan Maury: Wow. I mean, that's not super surprising to hear about, you know, sort of current data I guess. But I imagine that wasn't always the case, right? Could you talk a little bit more about that and tell us like how reliable birth records are now?

>> Eric Jensen: Yeah. Next slide please. So you're right, Meghan, that hasn't always been the case. So, for instance, here's a picture of a Census Bureau report from 1947, and this discusses how the adjusted for underregistration produce accurate population estimates. Next slide please. Now here's a picture of a blank infant card that was used in the 1950 Census. This is really interesting. So in 1950, the census was a lot different than it was in

2020 because all the data were collected by enumerators, so there was no option for folks to fill out, you know, a paper form and there's definitely no internet option. But the enumerator would fill out one of these infant cards for babies born between December 1949 and March of 1950. So the cards were then matched to birth certificates to try and determine the completeness of the birth registration system at that time. And a big thing this test showed is that there were gaps between the number of babies born and those that were being registered. So a similar study was conducted in the late 1960s, and that study found that by that time most births were being registered. And then today we assume that 100% of births are registered. So until the early 1990s, we just used aggregated birth records from the National Center for Health Statistics. This means that we got data for the total number of births by some limited geographic and demographic characteristics. But beginning in 1993, we started receiving microrecords from the National Center for Health Statistics. These microrecords included a lot of rich information about births, for instance, the birth year, the birth month, time of birth, birthplace, information about the parents' age, nativity, resident status, race and ethnicity, even information about the parents' education level, whether the child has siblings, some health information about the parents and the child, information about the labor and delivery. For our population estimates, we only use some of this information. So using microrecords, or records for each individual birth, gives us a lot of flexibility to process the birth estimates at different levels of geography. Also we don't have issues with suppressed data, and this means, you know, data that are restricted or have limited access because of privacy concerns. And this is because we have a special agreement with the National Center for Health Statistics to receive these files. Next slide please. So even though using the microrecords is better, there's still some challenges. For instance, the birth records only include the race and Hispanic origin of the father and mother and not the child. So to produce estimates of births by race and Hispanic origin, we use a process that we call Kidlink. Now this is different from the Social Security Kidlink file that you may have heard of or may be aware of. For our Kidlink process, we use data from the census to identify the reported race and Hispanic origin of the parents and the reported race and Hispanic origin of biological children living in the same household. We then calculate the proportion of children in a certain race or Hispanic origin group given the race and Hispanic origin group of the parents. We then use these proportions to assign these characteristics to aggregated birth records. We also have a challenge with the birth records because there's normally a two-year lag from when the birth occurs and when we get the final birth records from the National Center for Health Statistics. So we've developed methods to estimate the births during this period where we don't have the full NCHS data. For part of that, we get help from our state partners in what's called the Federal-State Cooperative for Population Estimates, and we work with them to get updated birth numbers or birth estimates for their state.

>> Meghan Maury: Oh, that's fascinating. So I'm with you so far. What about records on deaths, on that second part of the equation?

>> Eric Jensen: Sure. Next slide please. So we use data from the same source, the National Center for Health Statistics, to calculate the deaths part of our equation. So I won't talk, you know, in as much detail as I did with the births because the process is nearly identical. For decades we used aggregated deaths, and then starting in the early 1990s and since then

we've used microrecords to produce death components for the population estimates. So also like the births, having the microrecords gives us a lot more flexibility to produce the estimates of deaths at different geographic levels. And another big advantage is that the files we get from the National Center for Health Statistics do not have this suppressed data, and that's important because it lets us produce estimates for small counties or small populations whose data are usually suppressed in the publicly available files. So we have similar challenges when using -- next slide please. We have similar challenges when using the death records. For instance, the race and Hispanic origin information on the death certificate is often filled in by the mortician or funeral director who may or may not consult with the family when doing this. So this information may not be accurate. In fact, there's some evidence that some race groups, such as Native American, Alaska Natives, are underrepresented and underreported in the death records. Also Hispanics may be underreported in the death records. The big issue is there's a two-year lag that we have, so the data are not always current. So from the year of the death until we get those records from the National Center for Health Statistics, there's that two-year period.

>> Meghan Maury: Wow. Two years just feels like a really long time right now, especially with the COVID-19 pandemic. I think it makes me curious about how readily accessible and accurate are these data on death. So on the one hand it feels like we were all hearing kind of daily updates on the, you know, heartbreaking number of people who were passing away due to COVID, so it might stand to reason that those records really were readily accessible, especially during the pandemic. But then on the other hand, you know, the pandemic caused so many just real huge adjustments for how information flows. And I'm curious about just how did that all equate into accessibility for these records?

>> Eric Jensen: Yeah. Next slide please. So you're right. Not having current deaths was a big challenge for incorporating the impact of COVID-19 into the population estimates. One thing was that the National Center for Health Statistics started making available provisional daily updates of COVID-19 deaths and total deaths by week, by month, and by year for 2020 and for 2021. Provisional data were available by state, county, and demographic characteristics, so this was huge. We were able to use these. However, there were some limitations with these data. So, for example, the data were only available for 1,100 of the 3,142 counties in the United States. And even fewer counties were available for most race groups. Still, these data allowed us to make adjustments to account for COVID to the deaths in the most recent population estimates in ways we couldn't have just relying on those NCHS data that again we get and they usually lag by two years. Next slide please. So we also use administrative data among other sources to estimate migration, and we usually break migration into two parts: International migration and domestic migration. Next slide.

>> Meghan Maury: Got it. So we need to factor in people who are moving into and out of the country, you know, another set of administrative data that I imagine was probably also impacted by COVID-19, right?

>> Eric Jensen: Right. Next slide. So for the international migration components of the population estimates, we need to measure all of the inflows and outflows of both the native and the foreign-born populations to and from the United States. So historically we've used a lot of different data sources to do this, so data from the Department of State, Department

of Homeland Security, and other federal agencies. And we do this to try to estimate the flows of immigrants, of nonimmigrants, of temporary workers, of international students, and refugees and asylees. So our current method mostly uses data from the American Community Survey, but we had to use administrative data to account for the unique international migration patterns that resulted because of the COVID-19 pandemic. So to do this, we took visa issuance data from the Department of State. We also had refugee processing data from the Department of State. We used data on international students from the International Institute on International Education. And then, finally, we had some administrative data on asylum filings. So we used these data to adjust the total level of the foreign-born population coming into the United States in 2020 and in 2021. Next slide please. So we also use administrative data to estimate domestic migration within the United States. Since the early 1970s, we've used tax returns from the Internal Revenue Service to do this. We also use Medicare enrollment records to estimate domestic migration for the older population who may not file tax returns. So specifically we take two consecutive years of tax returns that have been linked using something known as the personal identification key, or PIK. If the address has changed between the two years, then we consider that the household has moved in the past year. And we do a similar process with the Medicare enrollment records where we identify movers by a change in address between two years. So using these data is a big benefit over other approaches. For example, we could try to use a survey to estimate domestic migration, but even, for example, the American Community Survey, which is, you know, our largest household survey, has a small sample compared to the number of tax records that we get. Each year there are about 200 million individual tax returns. The ACS data are also lagged, meaning that they're at least a year old before we can use them. And we use the IRS data in the year that they're filed, so it's much more current. And then similarly, we get administrative data on everyone enrolled in Medicare. So having these very comprehensive data sources helps us make more accurate domestic migration estimates. There are some challenges with this method though. For example, not everyone files tax returns, so there could be coverage issues in the IRS data. One solution to this is that we use the returns to calculate rates of in-movers and out-movers that we then apply to the population, so this can kind of help with the potential coverage issues. There can also be issues where we aren't able to match the right people because of the process for creating that PIK that uses a probability matching. And one last challenge is the IRS data, like other administrative data, were not really collected to estimate domestic migration. That's just how we're using them. So, for example, sometimes we find lots of returns coming from the same address and we realize that this is the address of the tax preparation business and not an individual filer.

>> Meghan Maury: Oh yeah, that all makes so much sense. But it just never occurred to me how big those data sources are, so how much kind of scope there is in what you're getting there. Now I want to leave time for questions. We've got some questions in the Q&A. And just a reminder to our listeners that we really are eager to answer your questions, so if you have other questions, please pop them into that Q&A feature. But, you know, I also would love to have a quick overview of how administrative data play a role in demographic analysis while we have you since DA, as we call it, demographic analysis, is often sort of talked about in tandem with decennial census and population estimates.

>> Eric Jensen: Yeah, definitely. Next slide please. So you're right. The demographic analysis is a program where we develop special population estimates that are used to evaluate the quality of the census. So it's one of two methods that the Census Bureau uses to estimate net coverage error in the census. The other method is the post-enumeration survey. And for demographic analysis, we use current and historical birth and death records, data on international migration and also Medicare enrollment records to produce national-level estimates of the population that are completely independent of the census that we're evaluating. So the DA estimates are different from our annual population estimates because they're not calculated using census data. We don't use census data as the base population. We actually start with the birth records back to 1945 and then we build a population estimate by accounting for deaths and international migration for each birth cohort up until April 1st, 2020. So the Census Bureau started using demographic analysis to estimate coverage error in the census of 1960 and has used this approach every decade since, and earlier this year we released the DA net coverage estimates for the 2020 Census by age and sex. Next slide please. So the DA and PES results are a really important way of evaluating the accuracy of the 2020 Census. And one of the key findings from the results of these programs is that young children aged zero to four were undercounted in the 2020 Census at a higher rate than other age groups. This is consistent with past censuses where we also have seen large undercounts for young children. So earlier this year, the Census Bureau formed the Young Children Working Group which includes analysts from across the agency, and the goal of this team is to improve the coverage of young children in the census and in demographic surveys and to improve data for this population. One example of how the Census Bureau is improving data for young children is the population estimates. For the most recent series of annual population estimates, we've incorporated age and sex data from the 2020 demographic analysis estimates into our April 1st, 2020 estimates base. So this is the first time that we've done something like this, and it makes the estimates for young children higher than if we just use the 2020 Census results. So this is actually a good example of how leveraging the strength of administrative records and administrative data has helped us improve the undercount of young children in the population estimates. Next slide. So finally I want to mention another way that we use administrative data in the population estimates. So far I've been talking about population estimates, but we also produce estimates of housing units. And to do this, we start with data from the census on housing units and then we account for new construction. Also we account for conversions from nonresidential to residential. We have to figure out a way to subtract housing unit loss and then we do something for manufactured homes. So just real quick, as part of the housing unit loss component, we get administrative data from the Federal Emergency Management Agency, or FEMA, and we use these data to account for housing units that were destroyed in a natural disaster such as a wildfire or a hurricane. So just to sum up, we use administrative data for just about all the components of the population estimates we produce every year. And those administrative data have some distinct benefits, especially for coverage of populations, like young children, for instance. But they also lack some of the complexity that we're able to gain from data where people self-report things like race and ethnicity in a more nuanced way. That's why it feels real important to me that we're using all the tools in our toolbox to collect data and produce statistics, especially on these communities that have historically been underrepresented in our data.

>> Meghan Maury: That's so helpful, such a great framing and so helpful for helping me understand I think. You know, it's so interesting to see how these processes are similar between population estimates, decennial census, and demographic analysis but also how they complement each other. Like we can see that -- and please tell me if I'm mirroring this backright or not, but administrative data for the population estimates help us get even better coverage on certain populations like young kids than we might be able to necessarily get from surveys or censuses. But sometimes the data are less nuanced because those administrative data sources just have less detail in them maybe than what we're able to get from someone's self-response in a survey. Also we know surveys' and censuses' administrative data can help us fill in the blanks with more accuracy than from other sources. So in some places they're less nuanced than what we can get from, you know, a self-response, but they may be more nuanced than what we could get from like a proxy response or response from a neighbor or other statistical methods like maybe imputation. And honestly there's just so much more we can do. I've heard folks at the Bureau talk a little bit about the research we're doing this decade to figure out additional ways to fold in administrative data into our programs to improve the accuracy of our estimates and statistics. And it's so interesting. I can't wait to keep following it. So thank you so much -- thanks both you and Tom so much for sharing that baseline information. We do have a few questions in the Q&A. And I know we've been able to answer some of them on the spot, but I'd love to ask a couple of them here too so that the whole audience can benefit from their responses. So, Tom, if you're willing to join us back on the stage, we'd love to have you. And the first couple of questions are actually I think for you. So one question that came in a little bit earlier in our discussion was how do you deal with mobile housing units that are, you know, maybe changing addresses?

>> Tom Mule: All right. So as part of our administrative data usage, we wanted to be using administrative data where we knew it was good information for us to use, also identified where were there places were I had limitations and we needed to do our traditional field collection. So with mobile homes, the issues with those is they don't necessarily have a city style address, they don't have an address that we can accurately mail to, so it raises questions about -- you know, as part of my first bucket of cases that I'm talking, could we use this information to be able to determine whether it's, you know, vacant or it's not a housing unit. And I think it's part of the issue with it being such a transient population. Those were examples of one where I think we did not necessarily use administrative records and relied more on our traditional data collection methods.

>> Meghan Maury: That's really helpful to understand. And I think it keeps bringing me back to that concept that we talked a little bit in the 101 session about we only use these data where they're useful. I mean, if the data don't provide you with helpful information, you're going to look to another source or another method to get that information. So it's really helpful. I saw another question in the Q&A that I found really interesting. The person talked a little bit about where records may not reflect someone's identity accurately. So they gave an example of, you know, if the Social Security Administration has your race or ethnicity recorded in a different way than you self-identify. How are we meant to understand the quality of those records? Have we done testing on that? How do we know

how frequently the Social Security Administration data matches up with how people see themselves?

>> Tom Mule: All right. For the 2020 Census, we did extensive research in looking at the administrative data that we had either from our past census but also from other sources and being able to, especially with race and Hispanic origin, compared it to what did we collect in the 2010 Census, what did we collect in American Community Survey. And for these same people, what information did we have with Social Security Administration or Medicare or other programs? What were they providing? And we needed to see very high agreement between what the other agency was providing and what we had based on our 2010 Census or our American Community Survey, but we know the questions that we're asking that we're trying to get a lot of the details that the American public is interested in. So we did a lot of work to make sure we had that agreement to be able to use it in the 2020 Census.

>> Meghan Maury: That makes sense. So part of what you mean when you say, you know, good quality administrative records. And I think that leads me to another question that we got in the Q&A a bit earlier which is how does the process differ for decennial census if you have good administrative records, good quality administrative records, whether that means good quality like we have a good data source or good quality on that particular person or household versus what happens if we don't have good quality administrative records. How does that shift?

>> Tom Mule: Right. Along those lines you can kind of contrast what instances might we have where we have one instance. We have an address where we think we have good administrative record information. So an example of that was, okay, we have the Smith family. And so John and Mary Smith and their son Bobby they filled out their taxes on April 15th. They sent those in. We saw that their W2s earlier in the year were sent to the same address. Our third-party source was also showing that it's showing the Smith family at 101 Main Street. When we were delivering our questionnaires and our letter in March, they were getting delivered, they were not coming back as undeliverable as addressed. And if we see those happening, instances where ACS estimates are saying, hey, there's low mobility, there's low vacancy, that Smith family instance is one where we're more likely to make that as a case where we think it's administrative record occupied. We have a roster and we can reduce contacts. An example of where we might have administrative record data but we're less confident in it. You know, let's say, you know, we have, you know, Bob Jones, you know, Bob Jones lives at 5 Broad Street and Bob goes and he files his taxes in early February because Bob is getting a refund. So he's filing as early as possible to get that. But then we end up seeing when we send our mailings to 5 Broad Street in March that they're coming back as undeliverable as addressed for those. We also see information in other sources that Bob may have moved to another address. So based on those, we have information that, you know, Bob Jones was associated with 5 Broad Street, but that instance is there where based on all of the information that we have, those are instances where we need to do our traditional data collection. If we can't determine it was vacant or a housing unit, we need to, you know, do our traditional nonresponse followup field operation.

>> Meghan Maury: Yeah, that makes sense. It's just so fascinating to me. Now we've gotten a couple of other questions in the Q&A that I think are a little bit similar to the question that we got about mobile homes, about how we count other communities that made it be less lower represented in records, for example, people experiencing homelessness or people who don't think of themselves as experiencing homelessness but they don't' have kind of a stable address. Like maybe they're living in a vehicle or they're moving around a lot. We probably wouldn't rely on administrative records for counting those folks because those administrative records just may not have good data on those communities, so we'd rely on our regular operations to count people who are living in more transitory spaces. I think it always just comes back to that theme of we use administrative data when they're useful. And so that might just not be a place where we're looking to that data. But I do want to touch on one more theme that's come in in the Q&A, which is can you talk a little bit about some of the challenges of the process of linking administrative data to the census or to other data sources.

>> Eric Jensen: Yeah. I can talk about that. So for the pop estimates data that we link to the census, the way we do that, and the Census Bureau uses this method for, you know, other data sets, but we use what's called a personal identification key, or a PIK. And one issue of that is it's a probability, so we don't have like -- it's not like we have Social Security numbers that we can link to different data sources and stuff, like that we know exactly that's the right person. It's a probability. So we have all these characteristics, and we're pretty sure that's the right person. But there could be the case that we're linking people that aren't the right people. And also what usually happens is there's not enough information, and so some populations there might not be enough information to have a PIK or that assignment, that special unique assignment. So it is a problem, and there's a panel next week that's going to be talking about some of the challenges of linking different administrative -- about strengths and weaknesses of administrative data. And so I think they'll get into that a lot more next week.

>> Meghan Maury: Yeah. I'm really excited for that panel even as a listener I gotta say [laughs]. I think it's going to be some really interesting information and I think it may touch on all of these kinds of questions that we got here today. I apologize that we don't have time to get to everything that everyone asked, but it's so wonderful to have so many people on this session. Eric and Tom, I just wanted to ask if there's any last parting words you want to share with our listeners before we sign off for the day.

>> Tom Mule: We thank people for attending today's session.

>> Eric Jensen: Yeah. Thank you very much.

>> Meghan Maury: Excellent. Wonderful. Well, I do hope folks get a chance to tune in for the last two sessions in our series on October 18th, we'll join you again from 2 to 3 PM to talk through strengths and weaknesses, as we just discussed. And then our final session will be on October 20th from 3 to 4 PM where we'll get a chance to talk about some future possibilities for using administrative records to improve Census Bureau surveys and estimates. So I really hope that you'll come back to join us for both of those and that wraps us up for today. Thank you.

>> Greg Pewett: This concludes today's webinar. Thank you for your participation. You may disconnect at this time.